[19]中华人民共和国国家知识产权局

[51]Int.Cl[7]
G06F 17/30

[12]发明专利申请公开说明书

[21]申请号 00104774.4

[43]公开日 2000年10月18日

[11]公开号 CN 1270361A

[22]申请日 2000.3.28 [21]申请号 00104774.4

[30]优先权

[32]1999.4.9 [33]US [31]09/288,724

[71]申请人 国际商业机器公司

地址 美国纽约

[72]发明人 霍梅沃恩·萨德莫哈姆德·贝基
阿兰·查尔斯·路易斯·特里特施勒
玛荷什·维斯万纳坦

[74]专利代理机构 中国国际贸易促进委员会专利商标事务所

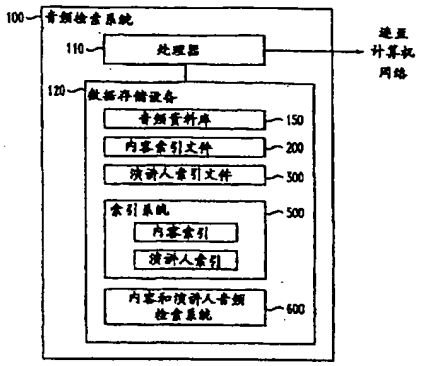
代理人 于 静

权利要求书4页 说明书17页 附图页数4页

[54]发明名称 使用内容和扬声器信息进行音频信息检索的方法和装置

[57]摘要

本发明公开一种根据音频内容和演讲人标识检索音频信息的方法和装置。基于 内容和基于演讲人的音频信息结果被结合在一起以提供对音频信息的引用。一个检索同包含一个文本串及一个给定的演讲人标识的文本查询相对应的信息的 查询搜索系统。一个对音频信息进行转换并建立索引以创建以时间标记的内容 索引文件和演讲人索引文件的索引系统。一个使用所产生的内容和演讲人索引 ,根据音频内容和演讲人标识执行查询－文档匹配的音频检索系统。



ISSN1008-4274

知识产权出版社出版

权　利　要　求　书

1. 一种从一或多个音频源检索音频信息的方法，所述方法包括步骤:

接收用户查询，该查询的约束至少要指定一个内容和一个演讲人；并且

将所述用户查询同所述音频源的一个内容索引和一个演讲人索引进行比较以识别满足用户查询要求的音频信息。

2. 根据权利要求 1 的方法，其中所述内容索引和所述演讲人索引是以时间进行标记的，并且所述比较步骤进一步包括在内容和演讲人领域对文档片段的起止时间进行比较的步骤。

3. 根据权利要求 1 的方法，其中所述内容索引包括所述音频源中每个词的出现频率。

4. 根据权利要求 1 的方法，其中所述内容索引包括所述音频源中每个词的倒排文档频率(IDF)。

5. 根据权利要求 1 的方法，其中所述内容索引包括所述音频源的长度。

6. 根据权利要求 1 的方法，其中所述内容索引包括一组链接指针，指向包含一个给定词的每个文档。

7. 根据权利要求 1 的方法，其中所述演讲人索引包括一个分值，指示一个已登记的演讲人模型到音频测试片段的距离。

8. 根据权利要求 1 的方法，其中所述演讲人索引包括每个音频片段的起止时间。

9. 根据权利要求 1 的方法，其中所述演讲人索引包括一个用来标识同片段相关的演讲人的标签。

10. 根据权利要求 1 的方法，其中所述的比较步骤进一步包括将满足基于内容查询的文档同满足基于演讲人查询的文档进行比较以识别相关文档的步骤。

11. 根据权利要求 1 的方法，进一步包括对所述音频源进行转换并建立索引以创建所述内容索引和所述演讲人索引的步骤。

12. 根据权利要求 11 的方法，其中所述创建所述演讲人索引的步骤包括在所述音频源中自动探测翻转并为每个所述翻转分配一个演讲人标签的步骤。

13. 根据权利要求 1 的方法，进一步包括将所述被识别出的音频信息的一部分返回给用户的步骤。

14. 根据权利要求 1 的方法，进一步包括给所述被识别出的音频信息的每个片段分配一个组合分值并将分级列表中所述被识别出的信息的至少一部分返回给用户的步骤。

15. 根据权利要求 14 的方法，其中所述组合分值评估了内容和演讲人两个领域之间的交迭程度。

16. 根据权利要求 14 的方法，其中所述组合分值得出了对基于内容的信息检索进行分级的分级文档分值。

17. 根据权利要求 14 的方法，其中所述组合分值得出了度量演讲人片段和登记演讲人信息之间接近程度的演讲人片段分值。

18. 根据权利要求 1 的方法，其中所述演讲人约束包括演讲人标识。

19. 根据权利要求 1 的方法，其中所述内容约束包括一或多个关键词。

20. 一种从一或多个音频源检索音频信息的音频检索系统，包括:

一个存储所述音频源的内容索引和演讲人索引以及计算机可读代码的存储器；及

一个同所述存储器连接工作的处理器，对所述处理器进行配置以实现所述计算机可读代码，所述计算机可读代码被配置用来:

接收指定一或多个词的用户查询并识别演讲人；及

结合基于内容和基于演讲人音频信息检索方法的结果，根据音频内容和演讲人标识提供所述音频源的引用。

21. 根据权利要求 20 的音频检索系统，其中所述内容索引和所述演讲人索引是以时间进行标记的，并且所述处理器进一步被配置用来在内容和演讲人领域对文档片段的起止时间进行比较。

22. 根据权利要求 20 的音频检索系统，其中所述内容索引包括所述音频源中每个词的出现频率。

23. 根据权利要求 20 的音频检索系统，其中所述内容索引包括所述音频源中每个词的倒排文档频率(IDF)。

24. 根据权利要求 20 的音频检索系统，其中所述演讲人索引包括一个分值，指示一个已登记的演讲人模型到音频测试片段的距离。

25. 根据权利要求 20 的音频检索系统，其中所述演讲人索引包括一个用来标识同片段相关的演讲人的标签。

26. 根据权利要求 20 的音频检索系统，其中所述处理器进一步配置用来将满足基于内容查询的文档和满足基于演讲人查询的文档进行比较以识别相关文档。

27. 根据权利要求 20 的音频检索系统，其中所述处理器进一步配置用来对所述音频源进行转换并建立索引以创建所述内容索引和所述演讲人索引。

28. 根据权利要求 20 的音频检索系统，其中所述处理器进一步配置用来给所述被识别出的音频信息的每个片段分配一个组合分值并将分级列表中所述被识别出的信息的至少一部分返回给用户。

29. 根据权利要求 29 的音频检索系统，其中所述组合分值评估了内容和演讲人两个领域之间的交迭程度。

30. 根据权利要求 29 的音频检索系统，其中所述组合分值得出了对基于内容的信息检索进行分级的分级文档分值。

31. 根据权利要求 29 的音频检索系统，其中所述组合分值得出了度量演讲人片段和登记演讲人信息之间接近程度的演讲人片段分值。

32. 一种从一或多个音频源检索音频信息的制造产品，包括:

一个其上包含计算机可读代码装置的计算机可读介质，所述计算机可读程序代码装置包括:

一个接收用户查询的步骤，该查询要指定一或多个词和一个演讲人的标识；及

一个结合基于内容和基于演讲人音频信息检索方法的结果，根据音频内容和演讲人标识提供所述音频源的引用的步骤。

33. 一种从一或多个音频源检索音频信息的制造产品，包括：

一个其上包含计算机可读代码装置的计算机可读介质，所述计算机可读程序代码装置包括：

一个接收用户查询的步骤，该查询约束至少要指定一个内容和一个演讲人；及

一个将所述用户查询同所述音频源的内容索引和演讲人索引进行比较以识别满足所述用户查询的音频信息的步骤。

说　　明　　书

使用内容和扬声器信息进行音频信息检索的方法和装置

本发明涉及信息检索系统，更确切地，涉及从一个多媒体数据库文件中检索满足用户指定要求的多媒体信息，如音频和视频信息的方法和装置。

信息检索系统主要集中在从大的文本集合中检索文本文档。文本检索的基本原理已经充分地提出并整理发布。例如，可参见G.Salton ,Automatic Text Processing, Addison-Wesley, 1989. 索引是一种将文档描述同查询描述进行匹配的机制。索引建立阶段(indexing phase)用一组字或词句对文档进行描述，而检索阶段(retrieval phase)用一组字或词句对查询进行描述。当文档描述同查询描述匹配时一个文档(或其中的一部分)得到检索。

多媒体对象，例如音频和视频文件所需的数据检索模型同文本文档所需的模型有很大的不同。对这些多媒体信息建立索引的标准特征集合有一点共性。对音频数据库建立索引的一种方法是使用某种音频提示，例如鼓掌，音乐或演讲。相似地，对视频信息建立索引的一种方法是使用关键帧，或相片的变化。对于有影响的演讲中的音频和视频信息，例如从广播中摘出的音频和视频信息，对应的文本可以使用语音识别系统得到，而转换文本可以用作建立相关音频(及视频)的索引。

当前的音频信息检索系统包含两个部分，即一个语音识别系统，用于将音频信息转换为用于建立索引的文本，和一个基于文本的信息检索系统。语音识别系统一般由三个部分组成，即词汇表，语言模型和一组针对词汇表中每个词的发音。词汇表是由语音识别器用来将语音翻译为文本的一组词。作为解码处理的一部分，该识别器将来自语音输入的声音同词汇表中的词进行匹配。因此，词汇表定义了可以被转换的词。如

果一个词不在词汇表中，则该词将得不到识别，不可识别的词必须首先被加入到词汇表中。

语言模型是同特定领域相关的词汇表中一系列词的数据库。其中还包括这些词以特定次序出现时的一组概率。当使用语音模型时，语音识别器的输出将偏向高概率词序。这样，正确的解码处理是判断用户所说的一系列词是否在语言模型中具有高概率。这样，当用户说了一个不常见的词序时，解码性能将下降。词的识别完全基于它的发音，也就是说，词的语音表示。为了得到最好的准确率，必须使用同特定领域相关的语言模型。建立这样一个语言模型需要明确的文本转换及语音。

基于文本的信息检索系统一般分两步进行工作。第一步是离线(off-line)建索引阶段，这时会收集同文本文档相关的统计信息来建立索引。第二步是在线(on-line)搜索并检索阶段，使用该索引来进行查询-文档匹配，随后将相关的文档(及附加信息)返回给用户。在建立索引阶段，会对语音识别系统的文本输出进行处理以得到在检索阶段用于快速搜索的文档描述。

在建立索引过程中，一般按序执行下列操作：(i)标记化(tokenization)，(ii)标记语音段落，(iii)形态(morphological)分析，及(iv)使用标准的结束词(stop-word)列表删除结束词。标记化探测语句边界。形态分析是一种语音信号处理的形式，它将名词分解为其词根，并附加一个指示复数形式的标记。同样，动词被分解为指示人，时态和语气的单元，并附加该动词的词根。关于 索引建立过程的一般性讨论可以参见于在此作为参照的 S.Dharanipragada et al., "Audio-Indexing for Broadcast News," in Proc. SDR97, 1997。

当用户使用这样一个基于内容的音频信息检索系统来检索其中包含一或多个在用户定义的查询中定义的关键词的音频文件时，当前的音频信息检索系统不允许用户根据演讲人标识有选择性地检索相关的音频文件。这样，需要一种方法和装置，可以根据演讲人标识和音频内容来检索音频信息。

一般而言，这里所揭示的是一种根据音频内容和演讲人标识用于检索音频信息的方法和装置。所揭示的音频检索系统将基于内容和基于演讲人的音频信息检索的结果结合在一起来提供对音频信息(并间接对视频)的引用。

根据本发明的一个方面，查询搜索系统检索同包含一个文本串(一或多个关键词)的文本查询及给定演讲人的标识相对应的信息。用户定义的查询约束(constraints) 同经索引的音频或视频数据库(或两者)进行比较并对包含与给定演讲人所说的指定词相关的音频/视频片段进行检索，展现给用户。

所揭示的音频检索系统由两个主要部分组成。一个检索系统，转换音频信息并对其建立索引以创建用时间标记的内容索引文件和演讲人索引文件；一个音频检索系统使用所生成的内容索引和演讲人索引，根据音频内容和演讲人标识执行查询-文档匹配。将相关的文档(及可能的附加信息)返回给用户。

通过比较内容和演讲人两个领域中文档片段的起止时间，对符合用户指定内容和演讲人约束的文档进行标识。根据本发明的另一个方面，内容和演讲人两个领域之间交迭的部分也已考虑在内。那些交迭较多的文档片段权重越高。通常，对于符合用户定义内容和演讲人约束的文档，使用下面的等式计算出一个组合分值分配给该文档:

组合分值=(分级文档分值+(lambda*演讲人片段分值))*交迭因子

分级文档分值对基于内容的信息检索进行分级，例如，使用 Okapi 等式。演讲人片段分值是一个距离度量值，用来指示演讲人片段和所登记的演讲人信息之间的接近程度，它可以在索引建立阶段进行计算。Lambda 是在对演讲人进行标识的过程中一个用于记录可信度的变量，它是一个介于 0 和 1 之间的值。

通常，交迭因子用来补偿完全没有交迭的片段，是一个介于 0 和 1 之间的值。根据本发明该组合分值可以用来对返回给用户的所标识的文档进行分级排序，将最匹配的片段放在列表的头部。

通过下面所参照的详细描述和附图，可以更完整地理解本发明以及本发明进一步的特征和优点。

图 1 是根据本发明的一个音频检索系统的方框图；

图 2A 是图 1 内容索引文件中文档数据库的一张表；

图 2B 是图 1 内容索引文件中文档存储块(chunk)索引的一张表；

图 2C 是图 1 内容索引文件中单字组（unigram）文件(词频)的一张表；

图 2D 是图 1 内容索引文件中倒排(inverse)文档索引(IDF)的一张表；

图 3 是图 1 中演讲人索引的一张表；

图 4 根据本发明示出了一个有代表性的演讲人的登记过程；

图 5 是一张流程图，描述了图 1 中音频检索系统所执行的一个示例性的索引建立系统过程；及

图 6 是一张流程图，描述了图 1 中音频检索系统所执行的一个示例性的内容和演讲人音频检索系统过程。

在图 1 中示出了根据本发明的一个音频检索系统 100。如下面所进一步讨论的，该音频检索系统 100 结合了两种根据音频内容以及演讲人标识来搜索音频资料以提供对音频信息(及间接对视频)引用的不同方法。特别地，用户指定的基于内容的检索结果，例如 Web 搜索引擎的结果，根据本发明将同基于演讲人的检索结果结合在一起。

本发明允许一个查询搜索系统检索同包含一个附加约束，也就是给定演讲人的标识的文本查询相对应的信息。这样，一个用户查询包括一个文本串，包含了一或多个关键词，以及给定演讲人的标识。本发明将用户定义查询的约束同一个经索引的音频及/或视频数据库进行比较，并检索相关的包含给定演讲人所说的指定词的音频/视频片段。

如图 1 所示，本发明的音频检索系统 100 包含两个主要部分，也就是说，一个转换音频信息并对其建立索引的音频检索系统 500，及一个音频检索系统 600。如下面所进一步讨论的，该索引建立系统 500 在索引建立阶段对语音识别系统的文本输出进行处理，建立内容索引和演讲

人索引。在检索阶段，内容和演讲人音频检索系统 600 使用索引建立阶段所生成的内容和演讲人索引，根据音频内容和演讲人标识进行查询-文档匹配，并将相关的文档(以及可能的附加信息)返回给用户。

如下面所讨论的，语音识别系统按每个词的时间顺序产生转换文本。同一般的信息检索场景不同，在转换文本中没有明显的文档，因此必须要人工生成。在所示的实施例中，对于基于内容的索引，同每个音频或视频文件对应的转换文本自动被划分为包含固定数量词，如 100 个词的交迭片段，并且将每个片段作为一个单独的文档来对待。在另一种实现方法中，使用标题识别模式将这些文件划分为多个标题。同样，对于基于演讲人的索引，音频或视频文件被自动划分为同给定演讲人相关的单独片段。这样，每当出现一个新演讲人讲话，就会产生一个新片段。

本发明通过基于内容的检索和基于演讲人的检索来确定音频，建立了音频的最佳部分。需要注意的是在基于内容的索引中，片段大小大约是讲 100 个词的时间，约 30 秒。但在基于演讲人的索引中，片段长度是可变的，它是演讲人变化探测器的一个函数。这样，不能预计片段长度。这样，根据本发明的特征，要同时对两个领域的片段起止时间进行比较。

根据本发明的一个进一步的特征，内容和演讲人领域之间交迭的部分也已考虑在内。那些交迭较多的文档片段权重越高。通常，如下面结合图 6 进一步讨论的，使用下面的等式计算出一个组合分值:

组合分值=(分级文档分值+(lambda*演讲人片段分值))*交迭因子

分级文档分值对基于内容的信息检索进行分级，例如，使用下面要讨论的 Okapi 等式。分级文档分值是一个查询项的函数，因此在检索时进行计算。演讲人片段分值是一个距离度量值，用来指示演讲人片段和所登记的演讲人信息之间的接近程度，它可以在索引建立阶段进行计算。Lambda 是在对演讲人进行标识的过程中一个用于记录可信度的变量，它是一个介于 0 和 1 之间的值。交迭因子用来补偿完全没有交迭的片段，是一个介于 0 和 1 之间的值。根据本发明该组合分值可以用来对

返回给用户的所标识的文档进行分级排序，将最匹配的片段放在列表的头部。

图 1 是一张方框图，示出了根据本发明的一个示例音频检索系统 100 的框架。音频检索系统 100 可以作为一个通用计算系统来进行实现，例如图 1 所示的通用计算系统。音频检索系统 100 包括一个处理器 110 和相关的存储器，如数据存储设备 120，它可以异地分布或放在本地。处理器 110 可以作为一个单独的处理器来进行实现，或者是几个本地或分布的以并行方式操作的处理器。数据存储设备 120 及/或一个只读存储器(ROM)用于存储一或多条指令，供处理器 110 来检索，解释并执行。

数据存储设备 120 最好包括一个音频资料数据库 150，用来存储一或多个根据本发明可以进行索引和检索的音频或视频文件(或两者都有)。另外，数据存储设备 120 包括一或多个内容索引文件 200 和一或多个演讲人索引文件 300，下面会结合图 2 和 3 分别进行讨论。通常，如下面结合图 2A 到 2D 所讨论的，内容索引文件 200 包括一个文档数据库 210(图 2A)，一个文档存储块索引 240(图 2B)，一个单字组文件(词频)260(图 2C)以及一个倒排文档索引(IDF)275(图 2D)。内容索引文件 200 及附加的索引信息在索引建立阶段借助语音识别系统生成，它将音频(或视频)文档描述为一组词或句的列表。演讲人索引文件 300 在索引建立阶段借助演讲人标识系统生成，并为一个音频文件每个片段提供一个演讲人标签。随后，在检索阶段，对内容索引文件 200 和演讲人索引文件 300 进行访问，如果内容索引文件 200 中的文档描述同用说指定查询的描述匹配并且由演讲人索引文件 300 中演讲人标签所指定的演讲人标识同指定的演讲人标识匹配，则检索一个文档。

另外，数据存储设备 120 包括程序代码，该程序代码将处理器 110 作为下面将结合图 5 进一步讨论的索引建立系统 500 和下面将结合图 6 进一步讨论的内容和演讲人音频检索系统 600 进行配置。如前所示，索引建立系统 500 对音频资料数据库 150 中一或多个音频文件进行分析并生成相对应的内容索引文件 200 和演讲人索引文件 300。内容和演讲人

音频检索系统 600 根据用户指定的查询来访问内容索引文件 200 和演讲人索引文件 300，根据音频内容和演讲人标识执行查询-文档匹配，并将相关的文档返回给用户。

索引文件

如前所示，首先对示例音频进行转换，例如，使用一个语音识别系统来产生音频信息的一个文档版本。随后，索引建立系统 500 对音频文件的文本版本进行分析，产生相对应的内容索引文件 200 和演讲人索引文件 300.

如前所示，内容索引文件 200 包括一个文档数据库 210(图 2A)，一个文档存储块索引 240(图 2B)，一个单字组文件(词频)260(图 2C)以及一个倒排文档索引(IDF)275(图 2D)。通常，内容索引文件 200 及附加的索引信息以一组词或句的列表的方式存储了文档的描述信息。在所示实施例中，内容索引文件 200 在其他信息中记录了 Okapi 等式所需的统计信息。

文档数据库 210(图 2A)维护了多条记录，例如记录 211 到 214，每条记录都同所示实施例中一个不同的包含 100 个词的文档存储块相关。在一种实现方法中，在文档之间有 50 个词交迭。对于每个在域 220 所标识的文档存储块，文档数据库 210 在域 222 和 224 分别指定该存储块的起止时间，并在域 226 指定文档长度。最后，对于每个文档存储块，文档数据库 210 提供一个指针，该指针同对文档存储块进行索引的文档存储块索引 240 相对应。尽管在所示实施例中文档具有 100 个词的固定长度，但字节长度是不同的。如下面所讨论的，文档长度(以字节表示)用于规范化信息检索的分值。

文档存储块索引 240(图 2B)维护了多条记录，如记录 241 到 244，每条记录都同所对应的文档存储块中的一个不同的词相关。这样，在所示实现方法中，在每个文档存储块索引 240 中有 100 条记录。在域 250 中对于每个词串(来自文档存储块)进行了标识，文档存储块索引 240 在域 255 中指示了该词的开始时间。

单字组文件(词频)260(图 2C)同每个文档相关，并指示出了每个词在文档中的出现次数。单字组文件 260 维护了多条记录，例如记录 261 到 264，每条记录同在文档中出现的一个不同词相关。在域 265 对每个词串进行了标识，单字组文件 260 在域 270 指示出了某词在文档中的出现次数。

倒排文档索引 275(图 2D)指示出了在文档集合(音频资料库)中每个词的出现次数，在出现某词的所有文档中，用它对当前文档的相关性进行评级。倒排文档索引 275 维护了多条记录，例如记录 276 到 279，每条记录同词汇表中的一个不同词相关。在域 280 中用词汇标识符对每个词进行了标识，倒排文档索引 275 在域 285 中指示了词串，域 290 是倒排文档频率(IDF)，域 295 是出现某词的所有文档的列表。域 295 中的文档列表使得不用进行实际搜索就可以判断出某词是否出现在某个文档中。

如前所示，图 3 所示的演讲人索引文件 300 为一个音频文件的每个片段提供了一个演讲人标签。演讲人索引文件 300 维护了多条记录，例如记录 305 到 312，每条记录同一个音频文件的不同片段相关。每个语音片段同不同的演讲人相关。域 325 中标识了每个片段，演讲人索引文件 300 在域 330 中标识了相应的演讲人，域 335 是包含某片段的相应的音频或视频文件。另外，演讲人索引文件 300 还在域 340 和 345 中分别指示出了某片段(如从文件开始处的偏移量)的起止时间。演讲人索引文件 300 在域 350 中设置了一个分值，用来指示结合图 5 如下面所讨论的，演讲人片段和所登记的演讲人信息之间的接近程度。

演讲人登记处理

图 4 示出了一个已知的用于注册或登记演讲人的处理过程。如图 4 中所示，对每个已注册的演讲人，演讲人的名字将随同一个演讲人训练文件，例如一个脉冲代码调制(pulse-code modulated，PCM)文件提供给演讲人登记处理 410。演讲人登记处理 410 对演讲人训练文件进行分析，在演讲人数据库 420 中为每个演讲人创建一条记录。将演讲人的声音样本加入演讲人数据库 420 的处理被称之为登记。登记处理是离线进

行的，音频索引系统假设这样一个数据库包含了所有感兴趣的演讲人。对每个演讲人大约需要一分钟的音频，该音频来自包含多种语音条件的多个声道和麦克风。已登记演讲人的训练数据或数据库使用层次结构进行存储，以便为有效的识别和检索而对模型的访问进行优化。

建立索引处理

如前所示，在索引建立阶段，图 5 中所示的索引系统 500 对来自语音识别系统的输出文本进行处理，执行内容索引和演讲人索引的建立。如图 5 所示，内容索引和演讲人索引的建立是沿两条平行处理分支进行实现的，在步骤 510 到 535 执行内容索引的建立，在步骤 510 及 550 到 575 执行演讲人索引的建立。但值得注意的是，内容索引的建立和演讲人索引的建立可以顺序执行，这对于在该技术方面具有一般技巧的人而言是很明显的。

作为内容索引建立和演讲人索引建立的初始步骤，对数倒频谱特征（cepstral feature）在步骤 510 以所知的方式从音频文件中提取出来。通常，步骤 510 将音频文件的域改为频率域，减小动态范围并进行逆向转换以将信号返回到时间域。

建立内容索引

然后音频信息被提交给一个转换引擎，例如 ViaVioce 语音识别系统，它可以从纽约 Armonk 的 IBM 公司以商业方式得到，在步骤 515 产生一个经转换的文件，其中的词以时间进行标记。随后，在步骤 520，这些以时间标记的词被收集到具有固定长度，例如所示实施例中 100 个词的文档存储块中。

内容索引文件 200 所需的统计信息可以在步骤 530 从音频文件中抽取。如上所讨论的，索引建立操作包括： (i)标记化(tokenization)，(ii)标记语音段落，(iii)形态(morphological)分析，及(iv)使用标准的结束词(stop-word)列表删除结束词。标记化探测语句边界。形态分析是一种语音信号处理的形式，它将名词分解为其词根，并附加一个指示复数形式的标记。同样，动词被分解为指示人，时态和语气的单元，并附加该动词的词根。

在步骤 530，索引系统 500 通过 Okapi 等式得到所需的统计信息。对每个在音频域标识的词，获取下面的信息：词频(在给定文档中某词的出现次数)；倒排文档频率(IDF)(标识出现某词的文档数目)；文档长度(为了规范化)和一组指向包含某词的每个文档的链接指针(一个倒排索引)。

在步骤 530 中获得的信息在步骤 535 被存储在内容索引文件 200 中，或者如果已存在内容索引文件，则对信息进行更新。

演讲人索引

如下面进一步所讨论的，基于演讲人的信息检索系统包含两个部分：(1)一个语音变化探测系统(通常称为演讲人分割)，及(2)一个独立于演讲人，语言及文本的演讲人识别系统。为了实现演讲人标识处理的自动化，在步骤 550 必须对非同类语音部分的边界(翻转)进行探测。每个同类片段应该对应单个演讲人的语音。一旦描述好后，每个片段可以按特定的演讲人进行分类(假设演讲人识别系统所需的最小片段长度需求)。

在所示实施例的步骤 550 用于划分语音片段的模型选择标准(model-selection criterion)是众所周知的贝叶斯信息标准(Bayesian Information Criterion，BIC)。输入音频流可以在对数倒频谱空间上通过高斯处理来进行建模。BIC 是探测高斯处理的(演讲人)翻转时最可能采用的方法。模型标识存在的问题是从一组候选模型中选择一个来描述给定的数据集。假设从输入音频信号导出的帧(10ms)是独立的并且是一次高斯处理的结果。为了在帧 i，1<=i<N，之后在一个 N 特征向量的窗口中探测是否存在语音变化，可以建立两个模型。第一个模型通过一次高斯处理表示整个窗口，其特征在于它采用的方法和完全协方差(full ccvariance) $\{\mu, \Sigma\}$。第二个模型用第一次高斯$\{\mu_1, \Sigma_1\}$表示窗口的第一部分直到帧 i，用另一次高斯 $\{\mu_2, \Sigma_2\}$ 表示窗口的第二部分。这样判别式可以表示为：$\Delta BIC(i)=-R(i)+\lambda P$，其中

$$R(i)=\frac{N}{2}\log|\Sigma|-\frac{N_1}{2}\log|\Sigma_1|-\frac{N_2}{2}\log|\Sigma_2|$$

及$P=\frac{1}{2}\left(d+\frac{d(d+1)}{2}\right)$是对窗口的补偿，$N_1$= i 是窗口第一部分的帧的数目，$N_2$= (N-i) 是第二部分帧的数目； d 是帧的维(dimension)。因此，P反映了模型的复杂程度，$d+\frac{d(d+1)}{2}$是用来表示高斯的参数数目。

ΔBIC<0 意味着，要考虑补偿，将窗口划分为两个高斯的模型较之仅用一次高斯来表示整个窗口的模型，更加合适。因此 BIC 象是一个阀值似然(thresholded-likelihood)比率标准，其中的阀值并不是利用经验来进行调节，它是有理论基础的。这一标准是非常强壮的并且不需要任何前继培训。

在所示实现中，为了在不损失精确度的情况下加快速度，这里采用了 BIC 算法。所使用的特征向量是简单的使用 24 维的美对数倒频谱（melcepstra）帧。在这些向量上没有进行其他处理。该算法运行在逐个窗口的基础上，并且在每个窗口，对一些帧进行测试以检查它们是否是 BIC 规定的片段边界。如果没有发现片段边界(ΔBIC<0 为正数)，则窗口大小增加。否则，记录旧窗口的位置，该位置也对应新窗口的开始位置(使用原来的大小)。

下面描述 BIC 的详细实现步骤。出于明显的实用原因考虑，BIC 并不是为窗口中的每个帧来实施计算的。而是使用帧的分辨率 r，它将窗口划分为 M=N/r 个子片段。在(M-1)次 BIC 测试的结果中， 选择造成ΔBIC<0 负数值最大的一个。如果存在这样一个负数值，则将探测窗口重置为其最小尺寸，并以探测到的点用一个更好的分辨率进行精炼。这些精炼步骤提高了整个计算次数并且影响这一算法的速度性能。因此，这些应在特殊的用户环境，实时或离线地剪裁掉。

如果没有发现负数值，窗口大小使用下面的规则 $N_i=N_{i-1}+\Delta N_i$，从 N_{i-1}增加到 N_i帧，当还没有发现变化时 N_i也进行增长： $N_i-N_{i-1}= 2(N_{i-1}-N_{i-2})$ 。在语音信号的同类片段中这将加速算法运行。为了不增加错误率，ΔN_i有一个上限。当探测窗口太大时，BIC 的计算次数可以进一步减小。如果提供的子片段超过 M_{max}，仅有 M_{max-1}次 BIC 计算会被执行--跳过第一次。

在步骤 555 中，使用步骤 550 的结果分析在步骤 510 产生的特征并生成话音片段，它由单个演讲人的多个语音存储块组成。话音片段在步骤 560 提交给演讲人标识系统。关于演讲人标识系统的讨论，可参见如 H.S.M.Beigietal.，“IBM Model-Based and Fram-By-Fram Speaker-Recognition，”in Proc. Of Speaker Recognition and Its Commercial and Forensic Applications, Avignon, France(1998). 通常，演讲人标识系统将话音片段同演讲人数据库 420(图 4)相比较并发现“最接近的”演讲人。

演讲人标识系统有两种不同的实现方法，基于模型的方法和基于帧的方法，它们都各有优缺点。引擎既不依赖于文本也不依赖于语言，这方便了例如广播新闻的实时语音材料的索引建立。

演讲人标识 -- 基于模型的方法

为了为数据库中的众多演讲人建立一组训练模型，根据一个具有 M 帧的语音序列，模型 M_i 为第 i 个演讲人服务，用 d 维特征向量 $\{\vec{x}_t\}_{t=1\ldots M}$ 进行计算。这些模型根据它们的统计参数进行存储，例如，当选择了高斯分布的情况下，$\{\vec{\mu}_{i,j},\Sigma_{i,j},\bar{C}_{i,j}\}_{j=1\ldots n_i}$ 由平均向量(Mean vector)，协方差矩阵，和计数(Counts)组成。

通过使用在 H.S.M. Beigi et. Al，“"A Distance Measure Between Collections of Distributions and Its Application to Speaker Recognition，”Proc. ICASSP98, Seattle, WA, 1998 提出的距离度量，为了比较这两种模型，建立一个层次结构来设计一种具有多种不同能力的演讲人识别系统，包括演讲人识别(证实声明)，演讲人分类(分配一个演讲人)，演讲人校验(通过将标签同多个其特征同那些带标签的演讲人相匹配的演讲人进行比较第二次审查(pass)以确认分类)，以及演讲人聚类(clustering).

为演讲人识别而设计的距离度量可以用不同数目的分布 n_i 计算两个模型之间的可接受距离．根据两个演讲人的模型的参数表示对他们进行单独比较，可以避免带入其他特征，从而使对两个演讲人进行比较的任务的计算强度大大减小．但是，识别阶段这种距离度量方法的一个缺点

是在比较计算开始前不得不使用整个语音片段来建立测试人(声明者)的模型。逐帧方法减轻了这一问题。

演讲人标识 -- 逐帧方法

用 M_i 表示对应第 i 个登记演讲人的模型。 M_i 完全由参数集进行定义，$\{\vec{\mu}_{i,j}, \Sigma_{i,j}, P_{i,j}\}_{j=1,\ldots,n_i}$ 包含了平均向量，协方差矩阵，以及每个演讲人 i 的高斯混合模型(Gaussian Mixture Model，GMM)的 n_i 个部分的混合权重。这些模型使用包含一个 M 帧的语音序列的训练数据，以及如上一节所描述的 d 维特征向量，$\{\vec{f}_n\}_{n=1,\ldots,M}$ 来进行创建。如果演讲人的总体大小为 N_p，则模型空间的集合为 $\{M_i\}_{i=1,\ldots,N_p}$。基本目标是找到 i，以便 M_i 能最好地解释以一个 N 帧序列，$\{\vec{f}_n\}_{n=1,\ldots,N}$ 表示的测试数据，或者作出判断，没有任何模型可以正确地描述数据。下面基于帧的方法对距离度量的可能性进行加权计算，$d_{i,n}$ 用于决策:

$$d_{i,n} = -\log\left[\sum_{j=1}^{n_i} P_{i,j} P\left(\vec{f}_n \mid M_i \text{第} j \text{部分}\right)\right]$$

这里，使用一个规范表示，

$$p(\vec{f}_n \mid \cdot) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_{i,j}\right|^{1/2}} e^{\frac{1}{2}(\vec{f}_n - \vec{\mu}_{i,j})' \Sigma_{i,j}^{-1} (\vec{f}_n - \vec{\mu}_{i,j})}$$

来自测试数据的模型 M_i 的总距离 D_i 是总测试帧数上所有距离的和。

为了进行分类，选择距语音片段距离最小的模型。通过对该最小距离片段和背景模型进行比较，可以提供一种方法指示原来的模型中没有一个匹配得很好。可选择地，可以使用选举技术来计算总距离值。

为了进行校验，一组预先确定的组成带标签演讲人群体的成员得到扩充，加入了多个背景模型。通过使用这一集合作为模型空间，如果声明者的模型具有最小距离则测试数据通过测试进行校验；否则，拒绝测试数据。

由于语音帧必须保留用来计算演讲人之间的距离，在训练中不使用距离度量。因此训练完成后，使用上面所讨论的基于模型的技术方法。

用于基于演讲人的检索方法的索引文件是在步骤 565 通过在演讲人分类和校验的结果上进行第二次审查而建立起来的。如果在步骤 565 对演讲人标识进行校验，则演讲人标签在步骤 570 被分配给片段。

如前所示，每个分类结果都伴随着一个用于指示从原来已登记的演讲人模型到音频测试片段之间距离的分值，相对于所关注的音频片(audio clip)的开始时间的片段起止时间，和一个标签(在登记期间对所提供的演讲人命名)。另外，对于任何给定的音频片，将收集分配给同一(演讲人)标签的所有片段。接着它们按它们的分值进行排序并用具有最好分值的片段进行规范化。对由系统处理并加入索引的每个新音频片，所有带标签的片段再次进行排序并重新规范化。

在步骤 575 中这一信息被存储在一个演讲人索引文件 300 中，或者如果演讲人索引文件 300 已经存在，则更新信息。

检索处理

如前所示，在检索阶段，图 6 中所示的内容和演讲人音频检索系统 600 使用在索引建立阶段生成的内容和演讲人索引根据音频内容和演讲人标识来执行查询-文档匹配，并将相关文档(及可能的附加信息)返回给用户。通常，可以使用两个不同的，非交迭的模块完成检索，一个用于基于内容的检索，另一个用于基于演讲人的检索。由于这两个模块是完全独立的，因此可以使用线索或进程来并发运行程序。在所示实现中两个模块顺序执行。

在检索时，内容和演讲人音频检索系统 600 在步骤 610 和 20 装入建立索引所使用的同一词汇表，标记字典，形态表和标记表。适当的内容索引文件 200 和演讲人索引文件 300 在步骤 620 被装入存储器。在步骤 625 直到接收到一个查询后执行测试。

查询串在步骤 630 进行接收和处理。为了响应一个所接收到的查询，在步骤 635 对查询串和内容索引文件 200 进行比较以使用一个目标分级函数(分级文档分值)计算出最相关的文档。根据本发明(步骤 645)，在对这些文档进行分级时使用的分级文档分值被记录下来用于随后组合分值的计算。

下面是用于计算文档 d 和一个查询 q 之间的分级文档分值的 Okapi 公式的版本:

$$S(d,q)=\sum_{k=1}^{Q} c_q(q_k)\frac{c_d(q_k)}{\alpha_1+\alpha_2\frac{l_d}{l}+c_d(q_k)}idf(q_k)$$

这里，q_k是查询中的第 k 项，Q 是查询中的项数，$c_q(q_k)$和 $c_d(q_k)$分别是查询和文档中第 k 项的计数，l_d 是文档的长度，l 是文档集合中文档的平均长度，$i_df(q_k)$是项 q_k的倒排文档频率:

$$idf(q_k)=\log\left(\frac{N-n(q_k)+0.5}{n(q_k)+0.5}\right)$$

其中，N 是文档的总数，$n(q_k)$是包含项 q_k 的文档的数目。这样，倒排文档频率项有助那些在文档中出现较少的项。(对于单字组，α_1=0.5，α_2=0.5)。很清楚，idf 及上面评分函数中除同查询相关的项之外的多数元素都可以预先进行计算并进行评分。

每个查询都同集合中的所有文档进行匹配，并且根据上面所示的 Okapi 公式所计算出的评分对文档分级。经分级的文档分值考虑了每个查询项在文档中的出现次数，并结合文档长度进行了规范化。这一规范化处理消除了偏见，即一般喜爱较长的文档，因为较长的文档有更大的可能性包含较多的给定词。这一函数还有利于那些针对一个文档及很少出现在其他文档中的项。(如果使用第二次审查，可以使用第一次审查中的顶级文档作为训练数据，通过训练另一个文档模型对文档重新分级)

随后，在步骤 640 对所标识的文档(或其子集)进行分析以确定在演讲人索引文件 300 中所标识的演讲人是否匹配在查询中由用户指定的演讲人。特别地，满足基于内容查询的分级文档的时间范围将同那些满足基于演讲人查询的文档进行比较以使用交迭的起止时间来标识文档。演讲人检索中的一个单一片段可以同文本检索中的多个片段交迭。

在步骤 645 对任何交迭文档的组合评分按如下公式计算:

组合分值=(分级文档分值+(lambda*演讲人片段分值))*交迭因子

同上面所描述的方式。所有经评分的文档接着用得到匹配分 100 的最相关的文档进行分级并规范化。

通常，返回前 N 个文档给用户。这样，一组 N 个最匹配片段的起止时间，随同匹配分值，以及用于计算相关分值的被匹配词在步骤 650 返回给用户。每个组合结果的缺省时间同基于内容的搜索中相应文档的开始时间相同。(另一个选择是使用演讲人片段的开始时间。)结束时间被置为演讲人片段的结束(可以简单地让演讲人结束他的语句)。但是，出于可用性考虑，片段可以用固定时间进行截取，如 60 秒，也就是说，两倍于平均文档长度。

用户界面

所示用户界面可以显示由检索引擎返回的所有 N 个选择的相关信息，对于使用媒体处理器部分的进一步选择，通过使用 Java 媒体过滤器进行实现，以通过类似 VCR 的界面显示 MPEG-1 视频。Java 应用负责定位视频文件(如果 PC 连网它可以放在服务器上)，然后使用在检索时收集的信息来修饰结果，例如显示所检索出的文档，相关信息，例如媒体文件名，开始时间，结束时间，分级，规范化评分，一个包含所检索出的媒体文件的图形视图，高亮显示查询词(以及其他影响文档评级的因素) - 这些仅当采用基于内容的搜索方法时才出现，否则为了回放，高亮显示检索出的文档的显示部分。

前 N 个被检索出的项以压缩形式呈现给用户。这样用户可以可视地重审检索出的项以进行下一步动作。通常，它包括所有所收集的关于包含文档部分文字的被检索文档的信息。当选中一个被检索项来细听/看音频或视频时，一旦定位媒体文件则调用媒体处理器部分，在指定开始时间之前，解压缩流(如果需要的话)，然后用音频或视频的第一帧初始化媒体播放器。类似 VCR 的界面允许用户从头到尾“播放”所检索出的视频或者在任何结合点停止和前进。

本发明的方法可以对基于内容的音频信息检索进行进一步的改进。从语音识别输出导出的当前文档集可以通过包含识别器对每个词或句进行的次好猜测(next-best guesses)进行扩充。 可以用这一信息来对索引

项，查询扩充以及检索进行加权计算。另外，在仅使用纯语音建立索引并进行检索时，通过用音乐或主要噪音探测片段可以得到更准确的识别。当前音频索引方法的一个限制是在语音识别器中使用了有限的词汇表。从信息检索的角度认为是非常重要的适当名词及缩写经常在词汇表中找不到，因此在转换文本中也找不到。克服这一限制的一个方法是用一个针对词汇表外词汇的文字监视器来补充语音识别器的功能。但是，为了让方法实用化，该方法必须具有以比实时快许多倍的速度在大量语音中探测所说词的能力。

可以理解，这里所描述的实施例和其变化仅仅是示例本发明的原理，那些熟悉该技术的人们可以在不偏离本发明范围和精神的前提下对本发明的实现做出多种修改。

说 明 书 附 图

图 1

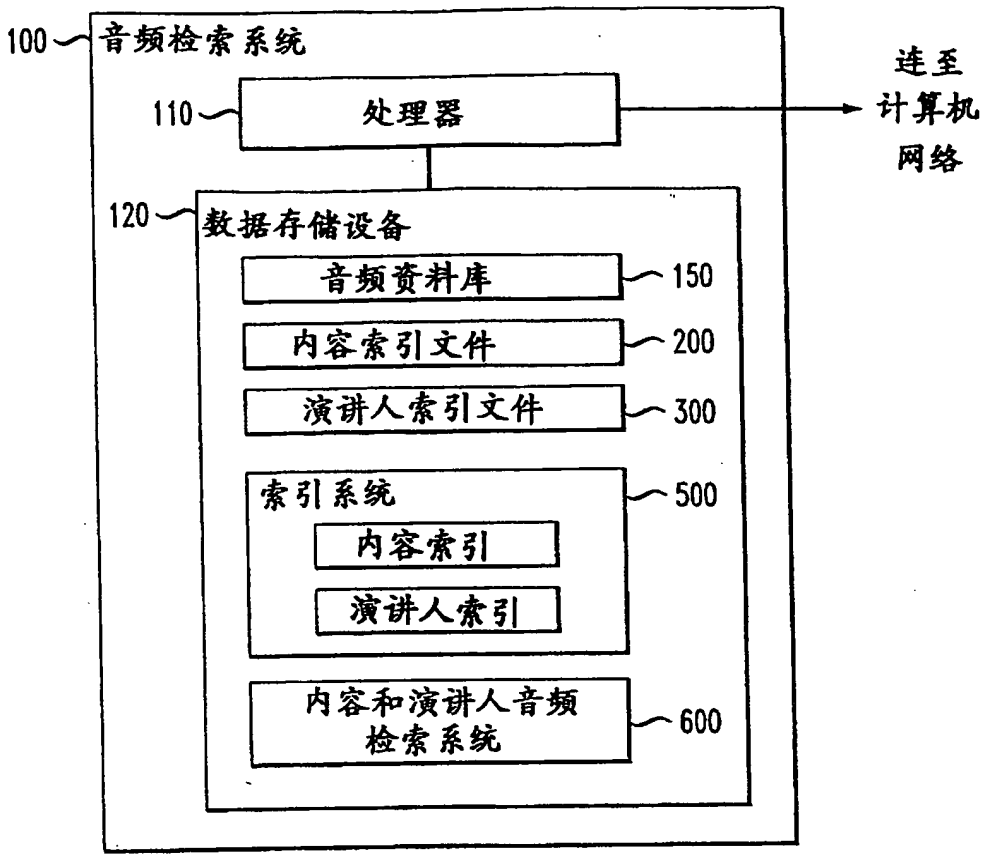


图 2A

文档数据库 210

	220 文档存储块标识	222 开始时间	224 结束时间	226 文档长度	228 文档存储块索引指针
211					
212					
213					
214					

图 2B

文档存储块索引
（文件索引块 N1）240

250	词串	开始时间	255
241	1	t_1	
242	2	t_2	
243	...	...	
244	N	t_N	

图 2C

单字组文件(词频)
260

265	词串	在文档中的出现次数	270
261	1	t_1	
262	2	t_2	
263	...	...	
264	N	t_N	

图 2D

倒排文档索引 275

280	285	290	295
词汇表标识	词串	IDF	文档列表
276			
277			
278			
279			

图 3

演讲人索引文件300

	325 片段编号	330 演讲人标签	335 音频标识符	340 开始时间	345 结束时间	350 分值
305	1	演讲人 1	介质 1	T_A	T_B	S_{10}
306	2	演讲人 1	介质 6	T_K	T_L	S_{11}
307	...	...	...	...	...	...
308	N	演讲人 1	介质 3	T_E	T_F	S_{12}
309	1	演讲人 N	介质 4	T_G	T_H	S_{20}
310	2	演讲人 N	介质 5	T_I	T_J	S_{21}
311	...	...	...	...	...	...
312	N	演讲人 N	介质 7	T_M	T_N	S_{22}

图 4

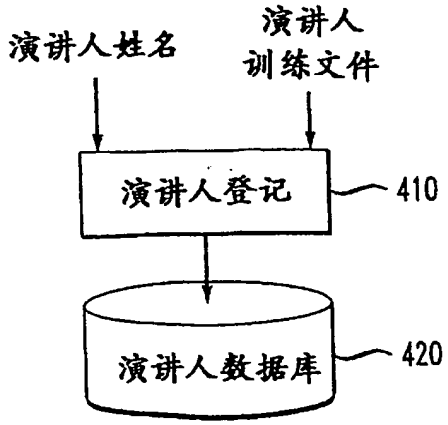


图 5

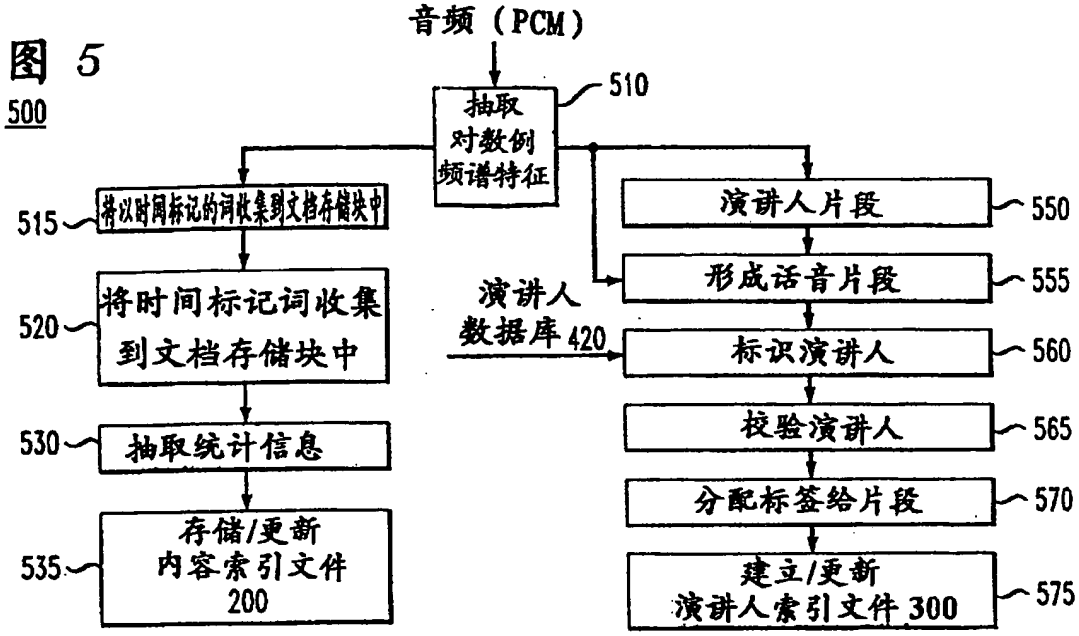
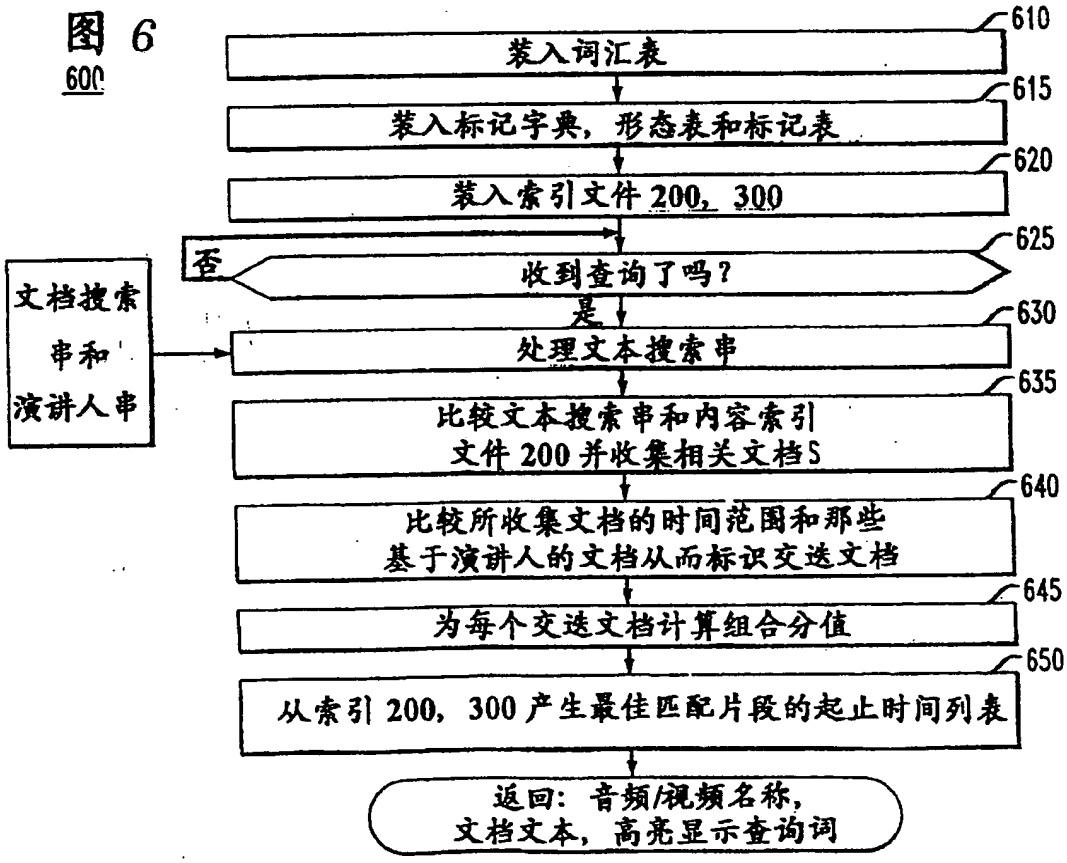

500
音频（PCM）
抽取对数例频谱特征
510
将以时间标记的词收集到文档存储块中
515
将时间标记词收集到文档存储块中
520
抽取统计信息
530
存储/更新内容索引文件200
535
演讲人片段
550
形成话音片段
555
演讲人数据库420
标识演讲人
560
校验演讲人
565
分配标签给片段
570
建立/更新演讲人索引文件300
575

图 6


600
装入词汇表
610
装入标记字典，形态表和标记表
615
装入索引文件200，300
620
否
收到查询了吗？
625
是
文档搜索串和演讲人串
处理文本搜索串
630
比较文本搜索串和内容索引文件200并收集相关文档S
635
比较所收集文档的时间范围和那些基于演讲人的文档从而标识交迭文档
640
为每个交迭文档计算组合分值
645
从索引200，300产生最佳匹配片段的起止时间列表
650
返回：音频/视频名称，文档文本，高亮显示查询词